Discrete Probability Density Estimation Using Multirate DSP Models

P. P. Vaidyanathan and Byung-Jun Yoon
Dept. of Electrical Engineering, California Institute of Technology
Pasadena, CA 91125, USA
ppvnath@systems.caltech.edu bjyoon@caltech.edu

Abstract. We propose a model based approach for estimation of probability mass functions for discrete random variables. The model is based on tools from multirate signal processing. Similar in principle to the kernel based methods, the approach takes advantage of well-known results from multirate signal processing theory. Similarities to and differences from wavelet based appoaches is also indicated where appropriate. In the final form, the probability estimates are obtained by filtering the square root of the histogram through a multirate system whose components are biorthogonal partners of each other.[1]

I. INTRODUCTION

The problem of estimating the probability density function (pdf) of a continuous random variable v from measurements has been of interest for many decades in the mathematics as well as signal processing communities. The most common method of using histograms and bar charts is quite satisfactory under some conditions. It has however been observed (e.g., see [1], [7]) that a model based approach has certain advantages especially when the number and quality of measurements is limited. For example, the "kernel" or model based method obtains an estimate of the pdf $f(v)$ by assuming that it has the form

$$f(v) = \sum_k c_k \phi(v - s_k) \tag{1}$$

where $\phi(v)$ is a fixed function (e.g., a spline, truncated Gaussian, etc). The preceding model seeks to represent the unknown pdf with a linear combination of shifted versions of the fixed function $\phi(v)$. With the shifts s_k typically fixed, one adjusts the coefficients c_k based on the measurements of the random variable v, so that the estimate $\widehat{f}(v)$ approximates the (unknown) pdf satisfactorily.

Figure 1(a) demonstrates the representation of $f(v)$ for the case where the shifts are uniform ($v_k = k\Delta$). Multimodal distributions can be approximated by choosing $\phi(v)$ to be a simple smooth function. With $\phi(v)$ chosen as a smooth function, the estimate $\widehat{f}(v)$ also enjoys the same smoothness. Figure 1(b) shows the special case where $\phi(v)$ is a discontinuous function (rectangular pulse) of duration Δ. In this case the estimate is similar to a histogram, especially when c_k is taken to be proportional to the number of measurements falling in the domain of the kth pulse $\phi(v - k\Delta)$. Further discussions on kernel based methods can be found in many references, e.g., [1], [7], [8].

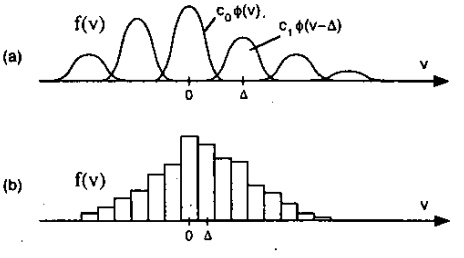

Figure 1. (a) The pdf estimate as a linear combination of shifted versions of a kernel $\phi(v)$, and (b) special case where $\phi(v)$ is the rectangular pulse as in histograms.

In this paper we consider model based estimation of pdf with the difference that the random variable be restricted to have uniformly spaced values (assumed to be integers with proper scaling). Thus the pdf $f(n)$ is a function of integer argument n. We present models based on multirate filters and filter banks and demonstrate their advanvtages. The analogy to continuous models such as (1), and the relation to wavelet methods [3] for pdf estimation will become clear in Sec. II as the theme evolves.

All notations are as in [4]. Thus $\downarrow M$ and $\uparrow M$ represent the M-fold decimator and expander respectively, and $[X(z)]_{\downarrow M}$ represents the z transform of the decimated version $x(Mn)$. Similarly $[X(z)]_{\uparrow M} = X(z^M)$ (z-transform of the expanded version).

II. MULTIRATE FILTER MODELS FOR PDF

Let $f(n)$ be the pdf of an integer random variable n. We assume that $f(n)$ can be represented as the output of an M-fold interpolation filter as shown in Fig. 2. Both M and $G(z)$ are assumed to be fixed. The input signal $c(k)$ is the free parameter to be adjusted based on measurements of the random variable n. We have

$$f(n) = \sum_k c(k) g(n - kM) \tag{2}$$

which is a linear combination of $g(n), g(n \pm M), g(n \pm 2M)$, etc. Notice the analogy to the continuous case (1). Assuming that $c(k)$ and $g(n)$ are in ℓ_2, the pdf $f(n)$ is also in ℓ_2. It belongs to the subspace $V_0 \subset \ell_2$ spanned by the shifted impulse responses

$$\eta_k(n) \triangleq g(n - kM).$$

[1]Work supported in part by the ONR grant N00014-99-1-1002, USA.

0-7803-7663-3/03/$17.00 ©2003 IEEE

This paper was originally published in the Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, & Signal Processing, April 6-10, 2003, Hong Kong (cancelled). Reprinted with permission.

Since the interpolation filter of Fig. 2 can be considered as one channel of a filter bank, we need M such filters to make up the whole space ℓ_2. Thus V_0 is a proper subspace of ℓ_2. If $G(z)$ is a lowpass filter, we can regard V_0 as a lowpass subspace, and $f(n)$ is smooth. We can add one or more fine-scale compoments by adding more channels in the model as demonstrated in Fig. 3 with an example. In this paper we restrict attention to the single channel model of Fig. 2, and explain how exactly the estimation procedure works starting from measurements.

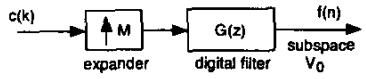

Figure 2. Definition of the subspace to which the pdf estimates are restricted.

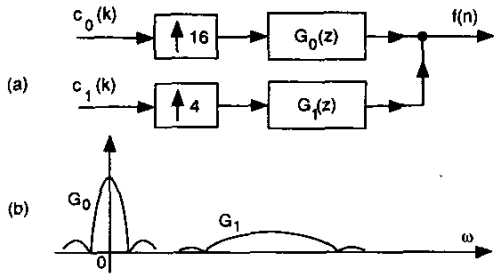

Figure 3. (a) Two channel model for the pdf $f(n)$, and (b) typical filter responses.

If we have a multichannel model with M branches and expanders $\{m_k\}$ satisfying

$$\sum_{k=0}^{M-1} 1/m_k = 1$$

then that would correspond, in principle, to a discrete time version of the full wavelet methods developed by Donoho et. al [3]. In this case the space allowed for $f(n)$ is the entire space ℓ_2. In this sense the model degenerates completely, but the value of the wavelet based method lies in the fact that we can perform denoising in the subbands and get cleaner estimates. A good review of these methods can be found in [6]. The wavelet method has several desirable properties asymptotically as the number of scales tends to infinity [7]. The method has also been applied for the estimation of hidden Markov model (HMM) parameters [2]. The model based approach (e.g., Figs. 2, 3) has fewer channels and does not satisfy $\sum_k 1/m_k = 1$. So V_0 is a proper subspace of ℓ_2. As a result, the variance of the estimate is smaller (as we shall demonstrate), compared to methods that do not rely on a model.

III. ESTIMATION FROM MEASUREMENTS

Consider again the model shown in Fig. 2 where $G(z)$ is a digitial filter, possibly FIR, more generally IIR. Suppose $H(z)$ is a filter such that

$$[H(z)G(z)]_{\downarrow M} = 1 \tag{3}$$

The filter $H(z)$ is said to be a biorthogonal partner of $G(z)$ with respect to M [5]. Note that $H(z) = 1/G(z)$ is a valid partner for any M but partners are not unique. It is shown in [5] that when $G(z)$ is FIR it is possible to find an FIR partner $H(z)$ under mild conditions on $G(z)$ (namely, the M polyphase components of $G(z)$ should not share a common factor).

In this paper, the importance of biorthogonal partners arises as follows: suppose a signal $f(n)$ can be represented as in Fig. 2 for some fixed $G(z)$, by appropriate choice of the driving signal $c(k)$. Then we can recover $c(k)$ from $f(n)$ by using the partner $H(z)$ as in Fig. 4.

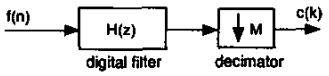

Figure 4. Reconstruction of the driving signal $c(k)$ from $f(n)$. Here $H(z)$ is the biorthogonal partner of $G(z)$.

To prove this note that the output of Fig. 4 has z-transform

$$\begin{aligned}[H(z)F(z)]_{\downarrow M} &= [H(z)C(z^M)G(z)]_{\downarrow M} \quad \text{(from Fig. 2)} \\ &= C(z)[H(z)G(z)]_{\downarrow M}\end{aligned}$$

Using (3) this reduces to $C(z)$ indeed. The preceding figure shows that $c(k)$ can be expressed in terms of the pdf $f(n)$ as follows:

$$c(k) = \sum_n f(n)h(Mk-n) \tag{4}$$

In this expression, the quantity n which is traditionally the "time index", should actually be interpreted as the integer random variable with pdf $f(n)$. The quantity $h(Mk-n)$ is a random variable because n is random. The right hand side of (4) is therefore the expectation of the random variable $h(Mk-n)$ with respect to the pdf $f(n)$. That is,

$$c(k) = E_f[h(Mk-n)] \tag{5}$$

The fact that expectations appear naturally in this manner is the crux behind nearly all model based and wavelet based pdf estimation methods [2,3,7]. Now assume that we have measurements $n_j, 0 \le j \le N-1$ of the integer random variable n. Then the expectation in (5) can be approximated by its sample-mean:

$$\widehat{c}(k) = \frac{1}{N}\sum_{i=0}^{N-1} h(Mk-n_i) = \sum_n \ell(n)h(Mk-n) \tag{6}$$

where $\ell(n)$ is the relative number of occurences of the integer value n in the measurements $\{n_j\}$. In other words, $\ell(n)$ is the histogram obtained from the measurements. Eq. (6) can be interpreted as a decimation filter operating on the histogram $\ell(n)$ as shown in Fig. 5. The quantity $\widehat{c}(k)$ is the estimate of $c(k)$ obtained from measurements.

The figure also shows how the estimate $\widehat{c}(k)$ is used in the model of Fig. 2 to obtain the pdf estimate $\widehat{f}(n)$.

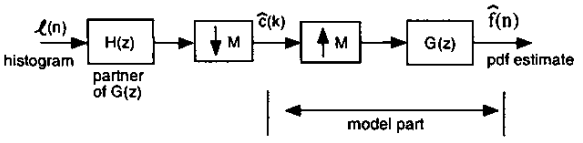

Figure 5. Estimation of $c(k)$ from the histogram $\ell(n)$, and subsequent estimation of the pdf $f(n)$.

Here then is the summary of how pdf estimation is done. We assume that the pdf $f(n)$ of the integer random variable satisfies the model of Fig. 2, i.e., belongs to the space V_0 (space of signals of the form (2)). We make measurements n_j of the random variable n, and compute the histogram $\ell(n)$, which represents an estimate of $f(n)$, but does not in general belong to the space V_0. We pass $\ell(n)$ through the system shown in Fig. 5 to obtain the refined estimate $\widehat{f}(n)$ which does belong to V_0. We can therefore regard the process as a projection of the histogram onto the space V_0 where $f(n)$ belongs.

Orthogonal projections. Given the histogram $\ell(n) \in \ell_2$, suppose we wish to find the signal $\widehat{f}(n) \in V_0$ closest to $\ell(n)$ in the least square sense, i.e., in the sense of minimizing

$$\|\ell(n) - \widehat{f}(n)\|^2 \triangleq \sum_n |\ell(n) - \widehat{f}(n)|^2$$

The result $\widehat{f}(n)$ is nothing but the orthogonal projection of $\ell(n)$ onto V_0. It can be shown [5] that if the filter $H(z)$ in Fig. 5 is such that

$$H(e^{j\omega}) = \frac{G^*(e^{j\omega})}{[|G(e^{j\omega})|^2]_{\downarrow M \uparrow M}} \tag{7}$$

then $\widehat{f}(n)$ is indeed the orthogonal projection of $\ell(n)$ onto V_0. Since $H(z)$ defined as above also happens to be a biorthogonal partner of $G(z)$ [5], it is said to be the least squares partner. With any other arbitrary partner $H(z)$, the projection $\widehat{f}(n)$ is "oblique" rather than orthogonal. The advantage of the orthogonal projection $\widehat{f}(n)$ is that it is guaranteed to be closer to the original pdf $f(n)$ than the histogram $\ell(n)$ is. That is,

$$\|\ell(n) - f(n)\| \geq \|\widehat{f}(n) - f(n)\| \tag{8}$$

To see this observe that since $\widehat{f}(n)$ is the orthogonal projection of $\ell(n)$ onto V_0, we can write $\ell(n) = \widehat{f}(n) + e(n)$ where $e(n)$ is in the orthogonal complement V_0^c of V_0. Thus

$$\ell(n) - f(n) = \widehat{f}(n) - f(n) + e(n)$$

Since $\widehat{f}(n) - f(n) \in V_0$ and $e(n) \in V_0^c$ it follows that

$$\|\ell(n) - f(n)\|^2 = \|\widehat{f}(n) - f(n)\|^2 + \|e(n)\|^2$$

proving (8). If the model filter $G(z)$ is chosen such that its magnitude square is Nyquist(M), that is,

$$\left[|G(e^{j\omega})|^2\right]_{\downarrow M} = 1 \tag{9}$$

then we have $H(e^{j\omega}) = G^*(e^{j\omega})$, i.e., $h(n) = g^*(-n)$. This is the most convenient choice in practice. The condition (9) is equivalent to the statement that the basis $\{g(n - kM)\}$ spanning V_0 is orthonormal. The filter $G(e^{j\omega})$ satisfying the orthonormality constraint (9) can be designed using one of many possible methods [4].

IV. THE SQUARE ROOT MODEL

The pdf model shown in Fig. 2 suffers from one disadvantage. Namely, for arbitrary input $c(k)$, the positivity condition $f(n) \geq 0$ may not be satisfied. This is important in the estimation process where Fig. 5 is used. There is no assurance that $\widehat{f}(n)$ will remain nonnegative. In fact if we design the filter $G(e^{j\omega})$ such that Eq. (9) holds, then $g(n)$ has some nonnegative coefficients and so does $h(n) = g^*(-n)$ (unless $G(z)$ has order $\leq M$). The estimate $\widehat{f}(n)$ is most likely to have some negative coefficients because of this.

A simple way to overcome this is to use the model output to represent a square root $f_s(n)$ of the pdf $f(n)$ rather than the pdf itself (Fig. 6(a)). We can still use Fig. 5 for the estimation of the pdf with slight procedural modifications as follows. First compute the histogram $\ell(n)$ from measurements as before, and take the square root signal $\sqrt{\ell(n)}$ as the input signal in Fig. 5. This is shown in Fig. 6(b). The final output, indicated as $\widehat{f}_s(n)$ is squared to get the pdf estimate $\widehat{f}(n) = [\widehat{f}_s(n)]^2$ which is therefore guaranteed nonnegative, and can be scaled to satisfy $\sum_n \widehat{f}(n) = 1$. Note that the intermediate signal $\widehat{c}_s(k)$ in the figure does not have the significance of an expectation, as did $c(k)$ (see Eq. (5)). However the method yields a very good (and nonnegative!) pdf estimate as we shall see. The value of the method in Sec. III is that it gives rise to the square root method logically.

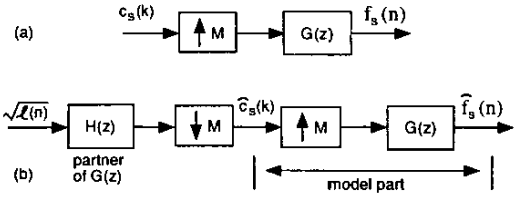

Figure 6. (a) The square root model, and (b) estimation of probability density using the square root method.

As in Sec. III the filter $G(z)$ can be designed to satisfy the orthonormal basis constraint (9) and $H(e^{j\omega})$ taken as the least squares biorthogonal partner $H(e^{j\omega}) = G^*(e^{j\omega})$. In this case $\widehat{f}_s(n)$ is the orthogonal projection of $\sqrt{\ell(n)}$ onto V_0 and we still have

$$\|\sqrt{\ell(n)} - f_s(n)\| \geq \|\widehat{f}_s(n) - f_s(n)\|$$

That is, as in Sec. III the projection can only bring the result closer to $f_s(n) = \sqrt{f(n)}$. One final crucial remark is in order. When we compute the square root $\sqrt{\ell(n)}$ we have the freedom to choose different signs for different n. Thus $\sqrt{\ell(n)} = s(n)\left|\sqrt{\ell(n)}\right|$ where $s(n)$ is a signature sequence (sequence of 1 and -1). We found experimentally, that the quality of the estimate $\widehat{f}_s^2(n)$ depends on the choice of $s(n)$. In the examples to follow, this has been chosen optimally by brute force search. More elegant ways to optimize $s(n)$ will be addressed in the future.

V. EXAMPLES AND CONCLUSIONS

To demonstrate the ideas we assume $M = 2$ and take $G(z)$ in Fig. 6(a) to be an FIR filter of order 13 such that the orthonormality condition (9) is satisfied. For this, we designed $G(z)$ as a spectral factor of a maximally flat halfband filter (Sec. 11.5.4 in [4] with $K = L = 7$). Notice that this $G(z)$ generates one of Daubechies' maximally regular orthonormal wavelets. The least squares biorthogonal partner is $H(e^{j\omega}) = G^*(e^{j\omega})$. By choosing the driving signal $c(k)$ appropriately, a signal $f_s(n)$ of length 18 was generated, and the test pdf $f(n) = f_s^2(n)$ obtained. The histogram $\ell(n)$ was generated using 100 measurements of the random variable drawn from the pdf $f(n)$.

Figure 7 (top) shows plots of the histogram and the pdf $f(n)$. Clearly the match is not very good. Next, we used the structure of Fig. 6(b) to generate $\widehat{f}_s(n)$. The quantity $\widehat{f}(n) = \widehat{f}_s^2(n)$, normalized for unit sum, is the model-based pdf estimate. Figure 7 (bottom) shows plots of the $\widehat{f}(n)$ and the pdf $f(n)$ showing that the agreement is excellent. The squared estimation errors are

$$\sum_n |f(n) - \ell(n)|^2 = 0.0082, \sum_n |f(n) - \widehat{f}(n)|^2 = 0.00078$$

Thus the error in the model based estimate is nearly 10 times smaller. Computation of the variance of the histogram estimate $\sigma_h^2(n) = var[\ell(n)]$ and the variance of the model based estimate $\sigma_m^2(n) = var[\widehat{f}(n)]$ (based on 100 experiments) showed that $\sigma_m^2(n)$ is much smaller for most n. In our experiment the total variances were

$$\sum_n \sigma_h^2(n) = 0.0082, \quad \sum_n \sigma_m^2(n) = 0.0038.$$

The reduced variance is due to the fact that $\widehat{f}_s(n)$ is restricted to a proper subspace $V_0 \in \ell_2$.

It is possible to modify the model based method to take care of measurement noise though this was not emphasized in this paper. As a concluding remark, notice that a practical challenge here is to identify the right subspace V_0 (i.e., the filter coefficients $g(n)$) that will be well suited for the class of pdf functions $f(n)$ of interest. This requires some apriori information on the source of $f(n)$. The optimization of the model filter $G(z)$ or filters $G_k(z)$ in a multichannel model, based on apriori information on the class of $f(n)$ of interest, is a topic for future work.

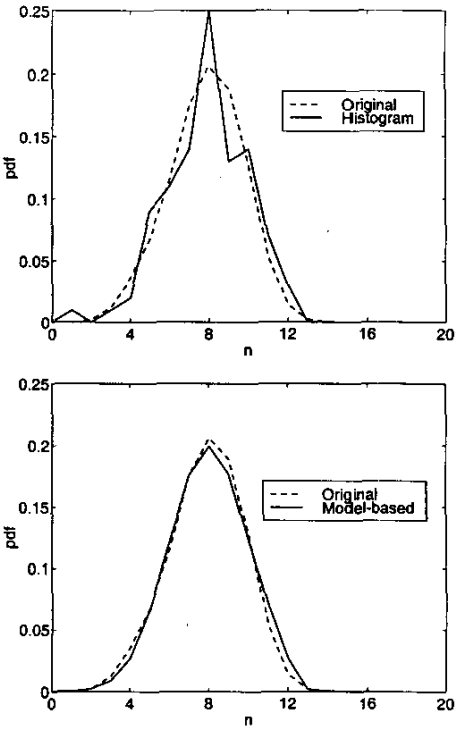

Figure 7. Top plot: the orginal pdf, and its estimate based on a histogram of measurements. Bottom plot: original pdf, and its estimate based on the multirate model.

REFERENCES

[1] Chiu, S. T. "Bandwidth selection for kernel density esitmation", Ann. Stat., vol. 19, pp. 1883–1905, 1991.

[2] Couvreur, L., and Couvreur, C. "Wavelet-based method for nonparametric estimation of HMMs", IEEE Signal Processing Letters, vol. 7, pp. 25–27, Feb. 2000.

[3] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picadrd, D. "Density estimation by wavelet thresholding", Ann. Stat., vol. 24, pp. 508-539, 1996.

[4] Vaidyanathan, P. P. *Multirate systems and filter banks,* Prentice Hall, Inc., 1993.

[5] P. P. Vaidyanathan and Bojan Vrcelj, "Biorthogonal partners and applications," IEEE Trans. Signal Processing, vol. 49, pp. 1013-1027, May 2001.

[6] Vidakovic, B. *Statistical modeling by wavelets,* John Wiley & Sons, Inc., NY, 1999.

[7] Walter, G. G., and Ghorai, J. K. "Advantages and disadvantages of density esitmation with wavelets", Comp. Sci., Stat., vol. 24, pp. 234–243, 1993.

[8] Walter, G. G. "Density esitmation in the presence of noise", Stat. and Prob. letters, 41, pp. 237–246, 1999.